Raw Sign and Magnitude Spectra for Multi-Head Acoustic Modelling

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Raw Sign and Magnitude Spectra for Multi-head Acoustic Modelling

Erfan Loweimi, Peter Bell and Steve Renals

Centre for Speech Technology Research (CSTR), School of Informatics, University of Edinburgh
{e.loweimi, peter.bell, s.renals}@ed.ac.uk

Abstract

In this paper we investigate the usefulness of the sign spectrum and its combination with the raw magnitude spectrum in acoustic modelling for automatic speech recognition (ASR). The sign spectrum is a sequence of ± 1s, capturing one bit of the phase spectrum. It encodes information overlooked by the magnitude spectrum enabling unique signal characterisation and reconstruction. In particular, we demonstrate it carries information related to the temporal structure of the signal as well as the speech's source component. Furthermore, we investigate the usefulness of combining it with the raw magnitude spectrum via multi-head CNNs at different fusion levels for ASR. While information-wise these two streams of information are together equivalent to the raw waveform signal the overall performance is noticeably higher than raw waveform and classic features such as MFCC and filterbank. This has been observed and verified in TIMIT, NTIMT, Aurora-4 and WSJ tasks and up to 14.5% relative WER reduction has been achieved.

Index Terms: Sign spectrum, raw magnitude spectrum, multi-head CNN, multi-stream processing, speech recognition

1. Introduction

Performance of speech recognition systems has dramatically improved over the last decade owing to the deep neural networks (DNNs), e.g. [1, 2]. DNNs essentially solve the data representation problem through learning a sequence of transforms which effectively filter out the task-irrelevant information and pass through the useful information w.r.t. the given task and objective function. Such information filtering, requires disentangling relevant and irrelevant information and minimising the effect of task-irrelevant input variabilities.

However, if task-relevant non-redundant information is lost during the feature extraction, then even a perfect pattern recognition system will not be able to compensate for it. Although the loss of task-irrelevant information within a pipeline can be helpful to a recognition system, underpinning such hand-engineered process where all and only useful pieces of information are passed through, is very difficult.

One solution to this problem is to bypass any initial lossy parameterisation stage and feed the DNN with the raw data. This approach, although providing the DNN with all signal information, substantially expands the input space and makes learning challenging. Nonetheless, recently raw waveform modelling has drawn much attention in the community and there is an expanding body of work in speech classification and recognition [3–13] that shows the effectiveness of this approach. Such models are more amenable to be interpreted but, generally speaking, in terms of performance are still lagging behind the systems with classic features. Ideally, we are interested in a representation which outperforms the classic handcrafted features while preserving all the signal information and bypassing any suboptimal information filtering occurs in feature engineering.

Another possible solution could be applying the full-resolution (or raw) magnitude spectrum which has been employed in ASR, e.g. in [14–17] and comparable results to classic features have been achieved. From a signal processing standpoint, a signal cannot be perfectly reconstructed from its magnitude spectrum [18, 19] which implies this spectrum does not carry all signal information. Therefore, even by using the entire magnitude spectrum already some information is discarded. However, Van Hove et al [19] demonstrated that the signal can be *perfectly* reconstructed when the magnitude spectrum is combined with the so-called *sign information* or sign spectrum.

In this paper, we propose a novel application of the sign spectrum along with the magnitude spectrum for speech recognition via multi-head CNNs. We study the usefulness of the sign spectrum and its information content via speech reconstruction experiments. Having shown it encodes information about the signal's temporal structure and speech's excitation component, we deploy it in acoustic modelling via a multi-head CNN system where the heads takes the raw magnitude and sign spectra. We also investigate the possible ways of fusing these two streams of information at different levels. Experimental results on TIMIT [20], NTIMIT [21], Aurora-4 [22] and WSJ [23] confirm that such representations which fully preserve the signal information, provide an improved performance compared to raw waveform models or classic features.

The rest of this paper is organised as follows. In Section 2, the definition, properties, information content and role of the sign spectrum in speech reconstruction are reviewed and scrutinised. Section 3 investigates the possible ways of combining the magnitude and sign spectra for optimal multi-stream information processing. In Section 4 the experimental results are presented and discussed. Section 5 concludes the paper.

2. Sign Spectrum

The Fourier Transform (FT) plays a central role in speech signal analysis. FT-based front-ends such as MFCC [24], PLP [25] and filterbank (FBank) are still widely employed at different tasks. These features are based on the magnitude spectrum of the FT and information-wise, include a lower amount of information than the raw magnitude spectrum due to the subsampling (approximately, average frequency pooling) done by the filterbank.

In fact, even if the entire magnitude spectrum, which hereafter we refer to as raw magnitude spectrum, is used, still some information is missed. In general, the signal cannot be uniquely specified by its magnitude spectrum [18]. This indicates the magnitude spectrum only contains a subset of the signal information and consequently a DNN trained with it, would be unable to exploit such information.

Speech is a mixed-phase signal owing to having a non-causal complex cepstrum [26]. Mixed-phase signals are decomposable into the minimum-phase and all-pass components [27].

Supported by EPSRC Project EP/R012180/1 (SpeechWave).

http://dx.doi.org/10.21437/Interspeech.2020-18

These two components are convolved in the time domain, multiplicative in the frequency domain and *orthogonal* in the information space. By orthogonality we mean, any knowledge about either one does not provide any useful prior information about the other one. That is, the raw magnitude spectrum contains all the information placed in the minimum-phase component but it does not contain the information in the all-pass part [28]. On the other hand, the all-pass part has a unit magnitude spectrum and its information solely resides in its phase spectrum.

If a DNN fed with both raw magnitude spectrum and the all-pass component, it has the opportunity to see and process all information. However, the all-pass component is complex, its magnitude is constant, devoid of any information and working with its principal phase is problematic due to the phase wrapping issue. On the other hand, if the inverse Fourier transform of the all-pass part is computed, although the wrapping issue is solved, the difficulties of raw waveform modelling arise. For example, the all-pass component is extracted via short-term (e.g. 25 ms) spectral processing while CNN-based raw waveform acoustic models require frames as long as hundreds of milliseconds (e.g. 200 ms [9, 11]) for optimal performance.

Another solution is the *sign* spectrum, proposed by Van Hove et al [19]. It complements the magnitude spectrum and does not have the difficulties of working with the all-pass part.

2.1. Definition

The sign spectrum, $S_X(\omega;\alpha)$, is defined as follows

$$S_X(\omega;\alpha) = \begin{cases} +1 & \alpha - \pi \le \phi_X(\omega) \le \alpha \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where $\phi_X(\omega)$ is the principal (wrapped) phase of the signal $x[n]$ and α is a constant in the range of $0 < \alpha \le \pi$. Fig. 1 depicts the sign spectrum along with the corresponding magnitude and phase spectra for a typical speech frame and $\alpha = \pi/2$. The sign spectrum for the frequency bins with phase spectrum in the green and red zones becomes $+1$ and -1, respectively.

With some algebraic manipulation, it can be shown that

$$S_X(\omega) = \text{sign}\{\text{Real}\{\exp(j(\frac{\pi}{2} - \alpha)X(\omega))\}\} \quad (2)$$

where sign is the $signum$[1] and $Real$ denotes the real part. The choice of the alpha does not affect the performance. It is typically set to $\pi/2$ which makes the sign spectrum equal to the algebraic sign of the $\text{Real}\{X(\omega)\}$ (Eq.(2)).

This sequence of ± 1s, is essentially one special bit of the phase spectrum information that is overlooked and missed by the magnitude spectrum ($|X(\omega)|$). Taking advantage of it, Van Hove et al defined the *signed-magnitude* spectrum, $\tilde{X}(\omega;\alpha)$,

$$\tilde{X}(\omega;\alpha) = |X(\omega)|\, S_X(\omega;\alpha) \quad (3)$$

and propounded the following theorem:

Theorem *Let $x[n]$ and $y[n]$ be two real, causal and finite extent sequences with z-transform which have no zeros on the unit circle. If $\tilde{X}(\omega;\alpha) = \tilde{Y}(\omega;\alpha)$ for all ω then $x[n] = y[n]$.*

From information standpoint, this theorem implies that the union of the magnitude and sign spectra information equals all signal information because by combining them, the signal is uniquely specifiable and recoverable. Also note that, these two components are orthogonal and do not share any information.

[1] With the exception that the algebraic sign of 0 is assumed to be 1.

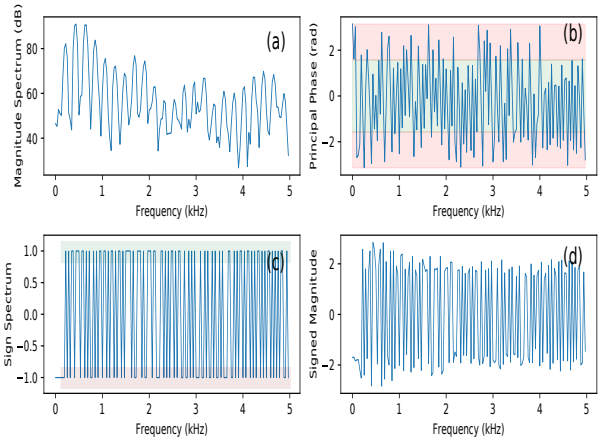

Figure 1: *Magnitude, phase, sign and signed-magnitude spectra for a typical speech frame. (a) Magnitude spectrum, (b) principal (wrapped) phase spectrum (assuming $\alpha = \pi/2$, green and red zone correspond to the bins where the sign spectrum takes 1 and -1, respectively), (c) sign spectrum for $\alpha = \pi/2$, (d) signed-magnitude spectrum. For a better visualisation, the signed-magnitude spectrum is compressed via* $\text{sign}(\tilde{X})|\tilde{X}|^{0.1}$.

2.2. Information Content of the Sign Spectrum

To illustrate the usefulness of the sign spectrum and its information content, we reconstruct the speech signal using the well-known Griffith-Lim [29] (GL) method in three modes: magnitude-only, sign-only and sign-magnitude-only signal. GL is an analysis-modification-synthesis (AMS) algorithm for iterative signal reconstruction from partial Fourier transform. The modification step for the sign-only and signed-magnitude-only reconstruction is done based on the proposed method by Van hove et al [19] (Eq.(28) in [19]).

To evaluate the quality of the reconstructed signals, we used all the 30 speech signals of the NOIZEUS [30] database and along with PESQ[2] [31] objective quality measurement score. PESQ typically returns values between 1 to 4.5 and the higher the better the quality. Number of iterations and frame overlap for GL reconstruction was set to 100 and 87.5%, respectively.

Fig. 2, depicts the spectrograms of the reconstructed signals when the signals are decomposed into 32 ms and 512 ms frames and Table 1 shows the corresponding PESQ scores. As seen in Fig. 2 (c) and (d), the sign-only reconstructed signal contains two pieces of information: first, in both short and long-term reconstruction, the events can be well localised in the time domain, contrary to the magnitude-only reconstructed signal in long-term, namely Fig. 2(b) in which the temporal structure in damaged. This implies sign information encodes timing information. Second, it contains the source (excitation) information.

It should be noted that although by frame length extension the importance of the sign spectrum increases, in short-term processing (32ms) it still could be helpful. As Table 1 shows, in this case it results in about 0.3 quality improvement in PESQ score which is a noteworthy gain.

Finally, the PESQ score of the Signed-magnitude-only reconstructed signal in 512 ms is not 4.5 (perfect quality). This may appear to contradict the aforementioned theorem. Actually, the theorem states what is possible in theory but does not determine the framework for realising it. While discussing the

[2] Perceptual Evaluation of Speech Quality

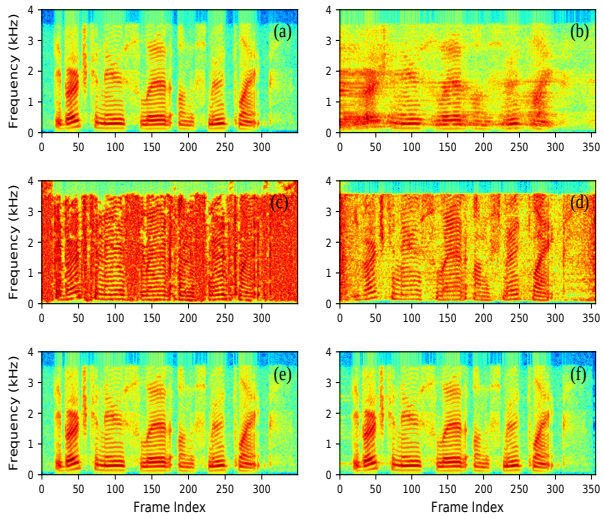

Figure 2: *Signal reconstruction using the Griffin-Lim method. Signal is decomposed into 32ms and 512ms frame lengths. (a) Magnitude-only, 32ms; (b) Magnitude-only, 512ms; (c) Sign-only, 32 ms, (d) Sign-only, 512 ms frames; (e) Signed-magnitude-only, 32ms; (f) Signed-magnitude-only, 512ms.*

Table 1: *PESQ score for the magnitude-only (Mag) and signed-magnitude-only (Mag+Sign) reconstructed signals via Griffin-Lim method for 32ms and 512ms frame lengths.*

	Hamming		Rectangular
	32 ms	512 ms	512 ms
Mag	4.22 ± 0.09	2.12 ± 0.24	2.38 ± 0.20
Mag+Sign	4.50 ± 0.00	4.20 ± 0.08	4.48 ± 0.02
Gain in PESQ	0.27	2.08	2.10

properties of the Griffin-Lim algorithm and the contributing parameters is outside the scope of this paper, as shown in Table 1, replacing the Hamming window with a rectangular window leads to near perfect reconstruction, in line with the theorem. For more about the window effect please refer to [26, 32, 33].

2.3. Statistical Properties of the Sign Spectrum

Having shown the role of the sign information in speech reconstruction, we wish to investigate its usefulness in speech recognition. Since statistical (mean and/or variance) normalisation (in utterance or speaker levels) is often applied as a helpful preprocessing step in DNN training [34], it is insightful to discuss the statistical properties of the sign spectrum and the necessity of normalising it. Regarding variance normalisation, it can be safely bypassed because the dynamic range of the sign spectrum is by definition limited to ± 1, and acoustical variabilities (noise, speaker, etc.) have no effect on its range. To evaluate its mean, we computed and compared its average with (log) magnitude spectrum at the utterance and speaker levels. As seen in Fig. 3, in both cases the mean is approximately zero for all bins. Therefore, the mean-normalisation, can be safely bypassed, too.

3. Combining the Magnitude and Sign Spectra via Multi-head CNN

As noticed, for speech synthesis the sign and magnitude spectra should be multiplied. However, for speech recognition and/or classification such a constraint may be relaxed and combination

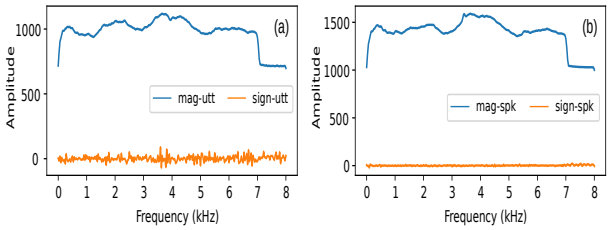

Figure 3: *Average magnitude and sign spectra at the (a) utterance (WSJ-eval92, utterance ID: 440c02010) level and (b) speaker (WSJ-eval92, speaker ID: 440) level.*

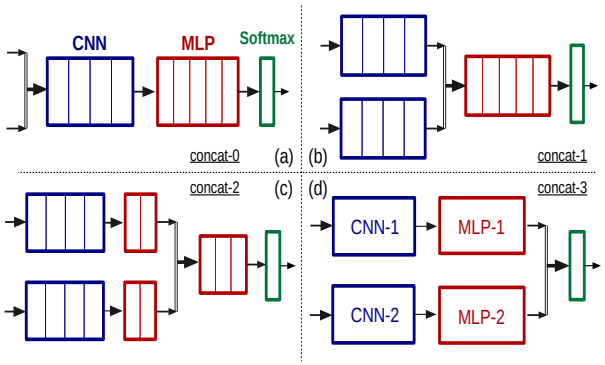

Figure 4: *Fusion (concatenation) of the magnitude and sign spectra information streams at different levels. (a) concat-0, (b) concat-1, (c) concat-2, (d) concat-3.*

can take any form in order to optimize the performance.

A direct combination, e.g. product, may not be optimal for classification because of the importance of these two disjoint pieces of information to the task, and also the way through which the information is encoded in each one is rather different (compare Fig. 1 (a) and (d)). These points necessitate dissociating the information processing workflow for these two orthogonal information streams. In other words, the chain of transforms to extract the task-optimal representation from these two signal elements should be different.

In this paper, we deploy a two-headed CNN for dealing with this multi-stream processing problem and investigate the optimality of information fusion at different levels. In our CNN-based framework, four levels for concatenation are considered: input level (**concat-0**); low level after the last convolutional layer (**concat-1**); medium level in the middle of the fully-connected (FC) layers (**concat-2**); and high level just before the softmax output layer (**concat-3**). Fig. 4 illustrates these fusion schemes for multi-stream processing.

Note that while the concat-1 to concat-3 fusion schemes are well-motivated choices, there is an issue with the concat-0, namely concatenation at the input level. Assuming the first layer is convolutional, the filters should learn task-relevant local correlations and relationships in the input. Such local patterns and consequently the corresponding matched filters are obviously different for the sign and magnitude spectra as they encode information in a different way (Fig. 1 (a) and (d)). As such using the same set of filters for both might perplex the learner.

4. Experimental Results

4.1. Setup

DNNs were trained using PyTorch-Kaldi [35, 36] with default recipes, including layer normalisation [37], batch normalisation [38] and dropout [39]. Monophone regularisation was removed.

The network consists of a cascade of four convolutional layers followed by an MLP with five fully-connected layers, each with 1024 ReLU units. Experiments where carried out on TIMIT, NTIMIT, WSJ and Aurora-4 (multi-style training). Alignments were taken from the respective Kaldi recipes [40]. For TIMIT and NTIMIT phone error rate (PER) and for WSJ and Aurora-4 word error rate (WER) is reported on standard development (Dev) and evaluation (Eval) sets. Aurora-4's test set consists of four subsets: A (clean), B (additive noise), C (channel mismatch) and D (additive noise plus channel mismatch). Ave in Table 3 is computed as follows: $(A + 6B + C + 6D)/14$. For computing the sign spectrum α was set to $\pi/2$. Feature normalisation for Aurora-4 was done on utterance level while for others done on speaker level. The sign spectrum was not statistically normalised, based on the discussion in Section 2.3. MFCC and FBank feature lengths are 39 (static+Δ+$\Delta\Delta$) and 80, respectively. Frame length for all features is 25 ms and each frame is augmented with ± 5 frames, except for raw waveform with 200ms frames and no context augmentation [9, 11].

4.2. Results and Discussion

Table 2 shows the PER of different features in TIMIT and NTIMIT tasks. The raw magnitude spectrum outperforms the classic features and its root compression ($\text{Mag}^{0.1}$) is helpful, in agreement with [15]. Based on this, we will use $\text{Mag}^{0.1}$ in all fusion experiments. Another interesting observation is that the sign spectrum alone can lead to a noteworthy and somehow surprising PER on both TIMIT (30.0%) and NTIMIT (54.6%) despite the fact that it is just one bit of phase information and merely a collection of ± 1s. As demonstrated in Section 2.2 and Fig. 3, sign spectrum captures information related to the temporal structure of the signal and speech excitation component.

Comparing Table 2 with Tables 3 and 4 shows that the fusion gain for TIMIT/NTIMIT is remarkably lower than Aurora-4 and WSJ. We believe this is owing to the fact that for effective distillation and combination of these two streams of non-redundant information, the DNN needs to see enough training data which may not be the case for database as small as TIMIT.

As seen in Table 3, for Aurora-4, WER reduction after fusing the magnitude and sign spectra is notably higher. Comparing the concat-1 fusion scheme with MFCC, filterbank and raw magnitude spectrum shows, respectively, 23.4%, 10.9% and 6.8% relative WER reduction which is a noteworthy gain. Similar to the phone recognition tasks, root compression of magnitude sightly helps and concat-1 outperforms concat-2 and concat-3. In addition, sign features alone lead to relatively good performance which encourages using this extra source of information overlooked by the magnitude-based front-ends.

For WSJ, as shown in Table 4, the same trend is observed: root compression of raw magnitude spectrum helps, sign spectrum alone returns a significant WER (14.0% on Eval-92) and concat-1 slightly outperforms other information fusion approaches. It returns 4.7% WER on Eval-92 without any data augmentation and relative WER reduction in comparison with the raw magnitude spectrum (5.5%) is 14.5%.

Why does concat-1 appear to be the optimal fusion scheme? Note that fusion at higher layers enlarges the network and increases model parameters (Fig. 4). This could run the risk of overfitting and might explain the higher performance of concat-1 as it has fewer parameters than others. However, this might be insufficient to draw a strong conclusion given the fact that many DNNs with high performance, operate in a remarkably overparameterised zone [41]. Another argument is that for a given fixed number of layers, the optimal fusion level should be high enough, to allow each information stream reaching a task-suitable level of abstraction and simultaneously, low enough, to leave sufficient layers on top to process the merged streams. Based on this argument, concat-1 provides the best trade-off.

Table 2: *TIMIT and NTIMIT PER for different front-ends.*

	TIMIT		NTIMIT	
	Dev	Eval	Dev	Eval
MFCC	17.1	18.6	27.5	28.9
FBank	16.3	18.2	27.5	28.5
Raw	17.2	18.6	25.2	26.3
Mag	16.8	17.8	30.9	30.1
$\text{Mag}^{0.1}$	15.9	17.6	25.2	25.6
Sign	27.2	30.0	53.7	54.7
Concat-1	15.4	17.5	24.3	24.8
Concat-2	15.7	17.8	24.8	25.3
Concat-3	15.5	17.5	24.6	25.6

Table 3: *Aurora-4 (multi-style) WER for different front-ends.*

Feature	A	B	C	D	Ave
MFCC	3.5	6.8	7.1	16.5	10.7
FBank	2.9	5.9	4.5	14.5	9.2
Raw	3.1	5.7	7.5	16.5	10.3
Mag	2.7	5.5	4.7	14.3	9.0
$\text{Mag}^{0.1}$	2.6	5.3	4.3	14.1	8.8
Sign	7.8	21.5	29.0	46.5	31.8
Concat-1	2.5	5.1	3.9	13.0	8.2
Concat-2	2.4	5.0	4.0	13.6	8.4
Concat-3	2.4	5.1	4.1	13.9	8.6

Table 4: *WSJ WER for different front-ends.*

	Dev93	Eval92
MFCC	10.4	6.8
FBank	9.1	5.9
Raw	8.4	5.2
Mag	9.3	5.9
$\text{Mag}^{0.1}$	8.8	5.5
Sign	21.2	14.0
Concat-1	8.1	4.7
Concat-2	8.2	4.8
Concat-3	8.2	4.8

5. Conclusions

In this paper, we investigated the usefulness of the so-called sign spectrum for ASR. The sign spectrum is a sequence of ± 1s encoding one bit of the phase spectrum information that complements the magnitude spectrum. That is, together with the magnitude spectrum it can perfectly characterise and reconstruct the signal. We scrutinised its information content through reconstructing the signal only from the sign spectrum and demonstrated it carries information related to the signal's temporal structure as well as speech's excitation component. Then, we studied the usefulness of fusing this overlooked stream of information with the raw magnitude spectrum via multi-head CNN. The fusion at low, medium and high levels has been explored in TIMIT, NTIMIT, Aurora-4 and WSJ tasks and notable performance gain was achieved. Applying such multi-stream processing framework which contains all signal information to other speech classification tasks is a broad avenue for future work.

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," *ArXiv e-prints*, 2016.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," *ArXiv e-prints*, 2017.

[3] N. Jaitly and G. E. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines." in *ICASSP*, 2011, pp. 5884–5887.

[4] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, pp. 1396–1407, 2011.

[5] D. Palaz, R. Collobert, and M. Magimai-Doss, "Analysis of CNN-based speech recognition system using raw speech as input," in *Interspeech*, 2015.

[6] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, 2015.

[7] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INTERSPEECH*, 2016.

[8] Z. Tuske, R. Schlüter, and H. Ney, "Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing," in *IEEE ICASSP*, 2018.

[9] M. Ravanelli and Y. Bengio, "Speaker and speech recognition from raw waveform with SincNet," in *ICASSP*, 2019.

[10] P. von Platen, C. Zhang, and P. C. Woodland, "Multi-span acoustic modelling using raw waveform signals," in *INTERSPEECH*, 2019.

[11] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[12] P. Noé, T. Parcollet, and M. Morchid, "CGCNN: Complex gabor convolutional neural network on raw speech," in *IEEE ICASSP*, 2020, pp. 7724–7728.

[13] T. Parcollet, M. Morchid, and G. Linarès, "E2E-SINCNET: Toward fully end-to-end speech recognition," in *IEEE ICASSP*, 2020, pp. 7714–7718.

[14] T. N. Sainath, B. Kingsbury, A. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *ASRU*, 2013.

[15] Z. Tuske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Interspeech*, 2014.

[16] Z. Zhu, J. H. Engel, and A. Hannun, "Learning multiscale features directly from waveforms," in *INTERSPEECH*, 2016.

[17] E. Cakir, E. C. Ozan, and T. Virtanen, "Filterbank learning for deep neural network based polyphonic sound event detection," in *IJCNN*, 2016, pp. 3399–3406.

[18] M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from phase or magnitude." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 6, pp. 672–680, 1980.

[19] P. Van Hove, M. Hayes, J. Lim, and A. Oppenheim, "Signal reconstruction from signed fourier transform magnitude," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 5, pp. 1286–1293, 1983.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[21] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *IEEE ICASSP*, 1990, pp. 109–112 vol.1.

[22] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep., 2002.

[23] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *IEEE ICASSP*, 1992, pp. 899–902.

[24] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 357–366, 1980.

[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[26] E. Loweimi, S. Ahadi, and H. Sheikhzadeh, "Phase-only speech reconstruction using very short frames." in *INTERSPEECH*. ISCA, 2011, pp. 2501–2504.

[27] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Prentice Hall, 2009.

[28] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain." in *INTERSPEECH*. ISCA, 2015, pp. 598–602.

[29] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 2, pp. 236–243, 1984.

[30] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE ICASSP*, 2001, pp. 749–752.

[32] D. Leigh and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578 – 616, 2007.

[33] E. Loweimi, S. Ahadi, T. Drugman, and S. Loveymi, "On the importance of pre-emphasis and window shape in phase-based speech recognition," in *Lecture Notes in Computer Science, Advances in Non-Linear Speech Processing (NOLISP)*, vol. 7911 LNAI, 2013, pp. 160–167.

[34] Y. LeCun, L. Bottou, G. B. Orr, and K. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade (2nd ed.)*, ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K. Müller, Eds. Springer, 2012, vol. 7700, pp. 9–48.

[35] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *IEEE ICASSP*, 2019.

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop on Autodiff*, 2017.

[37] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[41] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017.